

Recommender Systems with Generative Retrieval

Year: 2023 | Citations: 165 | Authors: Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan H. Keshavan, T. Vu

Abstract

Modern recommender systems perform large-scale retrieval by first embedding queries and item candidates in the same unified space, followed by approximate nearest neighbor search to select top candidates given a query embedding. In this paper, we propose a novel generative retrieval approach, where the retrieval model autoregressively decodes the identifiers of the target candidates. To that end, we create semantically meaningful tuple of codewords to serve as a Semantic ID for each item. Given Semantic IDs for items in a user session, a Transformer-based sequence-to-sequence model is trained to predict the Semantic ID of the next item that the user will interact with. To the best of our knowledge, this is the first Semantic ID-based generative model for recommendation tasks. We show that recommender systems trained with the proposed paradigm significantly outperform the current SOTA models on various datasets. In addition, we show that incorporating Semantic IDs into the sequence-to-sequence model enhances its ability to generalize, as evidenced by the improved retrieval performance observed for items with no prior interaction history.